Designation: D6708 – 24

An American National Standard

Standard Practice for Statistical Assessment and Improvement of Expected Agreement Between Two Test Methods that Purport to Measure the Same Property of a Material[1]

This standard is issued under the fixed designation D6708; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope*

1.1 This practice covers statistical methodology for assessing the expected agreement between two different standard test methods that purport to measure the same property of a material, and for the purpose of deciding if a simple linear bias correction can further improve the expected agreement. It is intended for use with results obtained from interlaboratory studies meeting the requirement of Practice D6300 or equivalent (for example, ISO 4259). The interlaboratory studies shall be conducted on at least ten materials in common that among them span the intersecting scopes of the test methods, and results shall be obtained from at least six laboratories using each method. Requirements in this practice shall be met in order for the assessment to be considered suitable for publication in either method, if such publication includes claim to have been carried out in compliance with this practice. Any such publication shall include mandatory information regarding certain details of the assessment outcome as specified in the Report section of this practice.

1.2 The statistical methodology is based on the premise that a bias correction will not be needed. In the absence of strong statistical evidence that a bias correction would result in better agreement between the two methods, a bias correction is not made. If a bias correction is required, then the *parsimony principle* is followed whereby a simple correction is to be favored over a more complex one.

NOTE 1—Failure to adhere to the parsimony principle generally results in models that are over-fitted and do not perform well in practice.

1.3 The bias corrections of this practice are limited to a constant correction, proportional correction, or a linear (proportional + constant) correction.

1.4 The bias-correction methods of this practice are method symmetric, in the sense that equivalent corrections are obtained regardless of which method is bias-corrected to match the other.

1.5 A methodology is presented for establishing the numerical limit (designated by this practice as the *between methods reproducibility*) that would be exceeded about 5 % of the time (one case in 20 in the long run) for the difference between two results where each result is obtained by a different operator in a different laboratory using different apparatus and each applying one of the two methods *X* and *Y* on identical material, where one of the methods has been appropriately bias-corrected in accordance with this practice, in the normal and correct operation of both test methods.

NOTE 2—In earlier versions of this standard practice, the term "cross-method reproducibility" was used in place of the term "between methods reproducibility." The change was made because the "between methods reproducibility" term is more intuitive and less confusing. It is important to note that these two terms are synonymous and interchangeable with one another, especially in cases where the "cross-method reproducibility" term was subsequently referenced by name in methods where a D6708 assessment was performed, before the change in terminology in this standard practice was adopted.

NOTE 3—Users are cautioned against applying the between methods reproducibility as calculated from this practice to materials that are significantly different in composition from those actually studied, as the ability of this practice to detect and address sample-specific biases (see 6.7) is dependent on the materials selected for the interlaboratory study. When sample-specific biases are present, the types and ranges of samples may need to be expanded significantly from the minimum of ten as specified in this practice in order to obtain a more comprehensive and reliable between methods reproducibility that adequately cover the range of sample-specific biases for different types of materials.

1.6 This practice is intended for test methods which measure quantitative (numerical) properties of petroleum or petroleum products.

1.7 The statistical calculations of this practice are also applicable for assessing the expected agreement between two different test methods that purport to measure the same property of a material using results that are not as described in 1.1, provided the results and associated statistics from each test method are obtained from a specifically designed multi-lab study or from a proficiency testing program (e.g.: ILCP) where

[1] This practice is under the jurisdiction of ASTM Committee D02 on Petroleum Products, Liquid Fuels, and Lubricants and is the direct responsibility of Subcommittee D02.94 on Coordinating Subcommittee on Quality Assurance and Statistics.

Current edition approved March 1, 2024. Published March 2024. Originally approved in 2001. Last previous edition approved in 2021 as D6708 – 21. DOI: 10.1520/D6708-24.

*A Summary of Changes section appears at the end of this standard

for each sample a single result is provided by each lab for each test method. The comparison sample set shall comprise at least ten different materials that span the intersecting scopes of the test methods with no material exceeding the leverage requirement in Practice D6300. Results and statistics shall meet requirements in 1.7.1. Requirements in this practice shall be met in order for the assessment to be considered suitable for publication in either method, if such publication includes claim to have been carried out in compliance with this practice. Any such publication shall include mandatory information regarding certain details of the assessment as specified in the Report section of this practice. R_{XY} shall be based on the published reproducibility of the methods.

1.7.1 For each test method and sample, results and statistics used to perform the assessment in 1.7 shall meet the following requirements:

(1) No. of results (N) ≥ 10,

(2) Anderson Darling statistic ≤ 1.12 (based on Normal Distribution),

(3) Standard Error (se_{sample}) is calculated using published reproducibility evaluated at the sample mean, N, and the factor 2.8 as follows:

$$se_{sample} = \left[R_{pub} / \left(2.8 \sqrt{N} \right) \right] \quad (1)$$

(4) se_{sample} is numerically less than [R_{pub} / (2.8 √10)], and

(5) Sample standard deviation (s_{sample}) per root-mean-square technique is not statistically greater than R_{pub} / 2.8 for at least 80 % of the samples in the comparison data set based on an F-test using 30 as the assumed degrees of freedom for R_{pub}, and (N − 1) for s_{sample} at the 0.05 significance level.

1.8 The methodology in this practice can also be used to perform linear regression analysis between two variables (X, Y) where there is known uncertainty in both variables that may or may not be constant over the regression range. The common acronym used to describe this type of linear regression is ReXY (Regression with errors in X and Y). The ReXY technique for assessing the correlation between two variables as described in this practice can be used for investigative applications where the strict data input requirement may not be met, but the outcome can still be useful for the intended application. Use of this practice for ReXY should be conducted under the tutelage of subject matter experts familiar with the statistical theory and techniques described in this practice, the methodologies associated with the production and collection of the results to be used for the regression analysis, and interpretation of assessment outcome relative to the intended application.

1.9 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D5580 Test Method for Determination of Benzene, Toluene, Ethylbenzene, *p/m*-Xylene, *o*-Xylene, C_9 and Heavier Aromatics, and Total Aromatics in Finished Gasoline by Gas Chromatography

D5769 Test Method for Determination of Benzene, Toluene, and Total Aromatics in Finished Gasolines by Gas Chromatography/Mass Spectrometry

D6299 Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance

D6300 Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products, Liquid Fuels, and Lubricants

D7372 Guide for Analysis and Interpretation of Proficiency Test Program Results

2.2 *ISO Standard:*[3]

ISO 4259 Petroleum Products—Determination and Application of Precision Data in Relation to Methods of Test

3. Terminology

3.1 *Definitions:*

3.1.1 *between ILCP method-averages reproducibility ($R_{ILCP\_\tilde{X},\ ILCP\_Y}$), n*—a quantitative expression of the random error associated with the difference between the bias-corrected ILCP average of method X versus the ILCP average of method Y from a Proficiency Testing program, when the method X has been assessed versus method Y, and an appropriate bias-correction has been applied to all method X results in accordance with this practice; it is defined as the numerical limit for the difference between two such averages that would be exceeded about 5 % of the time (one case in 20 in the long run).

3.1.2 *between-method bias, n*—a quantitative expression for the mathematical correction that can statistically improve the degree of agreement between the expected values of two test methods which purport to measure the same property.

3.1.3 *between methods reproducibility (R_{XY}), n*—a quantitative expression of the random error associated with the difference between two results obtained by different operators in different laboratories using different apparatus and applying the two methods *X* and *Y*, respectively, each obtaining a single result on an identical test sample, when the methods have been assessed and an appropriate bias-correction has been applied in accordance with this practice; it is defined as the numerical limit for the difference between two such single and independent results that would be exceeded about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of both test methods.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] Available from American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036.